Evaluating Hallucination in Large Vision-Language Models based on Context-Aware Object Similarities

Shounak Datta, Dhanasekar Sundararaman

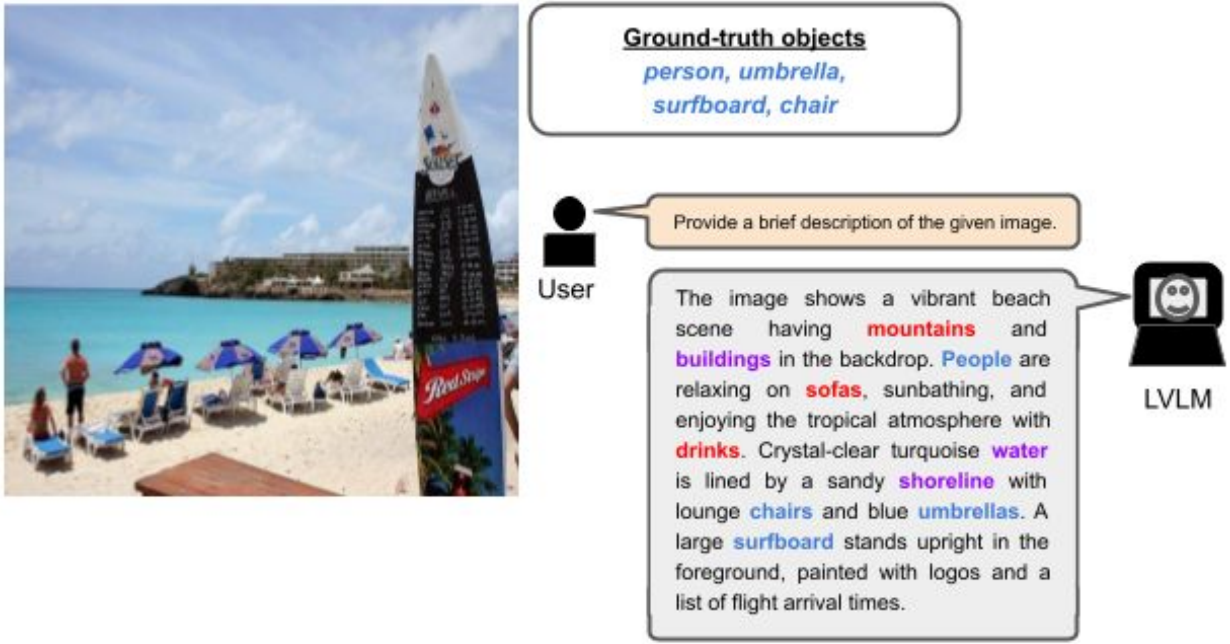
shounak.jaduniv@gmail.com, ds448@duke.edu

Authors

Shounak Datta

Dhanasekar (Dharun) Sundararaman

The problem of object hallucinations in LVLMs



Ground-truth objects

person, umbrella, surfboard, chair

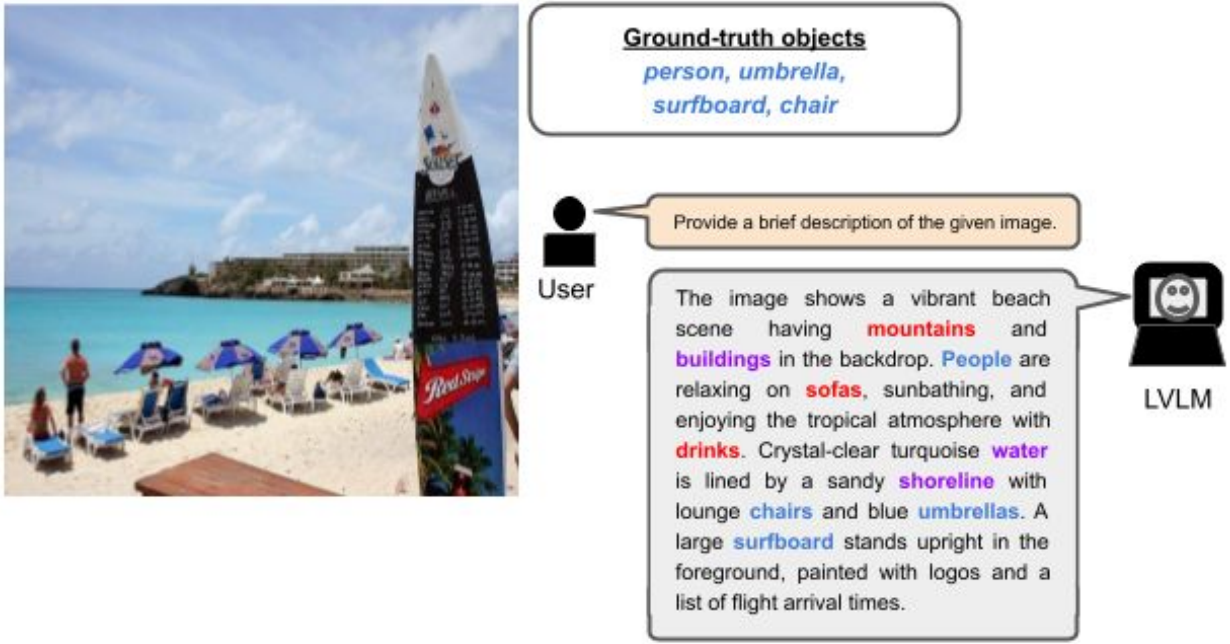
Provide a brief description of the given image.

The image shows a vibrant beach scene having **mountains** and **buildings** in the backdrop. **People** are relaxing on **sofas**, sunbathing, and enjoying the tropical atmosphere with **drinks**. Crystal-clear turquoise **water** is lined by a sandy **shoreline** with lounge **chairs** and blue **umbrellas**. A large **surfboard** stands upright in the foreground, painted with logos and a list of flight arrival times.

Object Hallucination

*Models generate descriptions of **objects** or entities **that do not exist** in the given visual input.*

The problem of object hallucinations in LVLMs



Provide a brief description of the given image.

The image shows a vibrant beach scene having mountains and buildings in the backdrop. People are relaxing on sofas, sunbathing, and enjoying the tropical atmosphere with drinks. Crystal-clear turquoise water is lined by a sandy shoreline with lounge chairs and blue umbrellas. A large surfboard stands upright in the foreground, painted with logos and a list of flight arrival times.

Issues

- Undermines the credibility of LVLMs despite their semantic coherence.
- Poses potential risks in real-world applications where accurate interpretation of visual content is important.

Existing work

- **Precision**
- **Recall**
- **CHAIR** (Rohrbach et al. 2018)
    - **CHAIR$_S$** - fraction of captions having hallucinations
    - **CHAIR$_I$** - avg. fraction of objects hallucinated (ie, 1 - Precision)
- **POPE** (Li et al. 2023)
    - *Claim:* Objects hallucinations are related to frequently occurring objects, and commonly co-occurring objects.
    - Probe the model with questions about the presence of objects in a given image:
        - randomly sample objects
        - top-k most frequent objects
        - top-k frequently co-occurring objects
    - "Independent" of generated captions.

Motivation

- Object hallucinations are known to be influenced by object statistics of the training data
    - Frequently occurring objects
    - Commonly co-occurring objects

Hypothesis: Due to autoregressive generation, previous objects mentioned in the generated part also influence the likelihood of object hallucinations in the remainder of the output?

For LLaVA-7B **20%** of the hallucinated objects co-occurred in the training dataset with at least one preceding objects.

Motivation

- Object hallucinations are known to be influenced by object statistics of the training data
  - Frequently occurring objects
  - Commonly co-occurring objects

  Hypothesis: Due to autoregressive generation, previous objects mentioned in the generated part also influence the likelihood of object hallucinations in the remainder of the output?

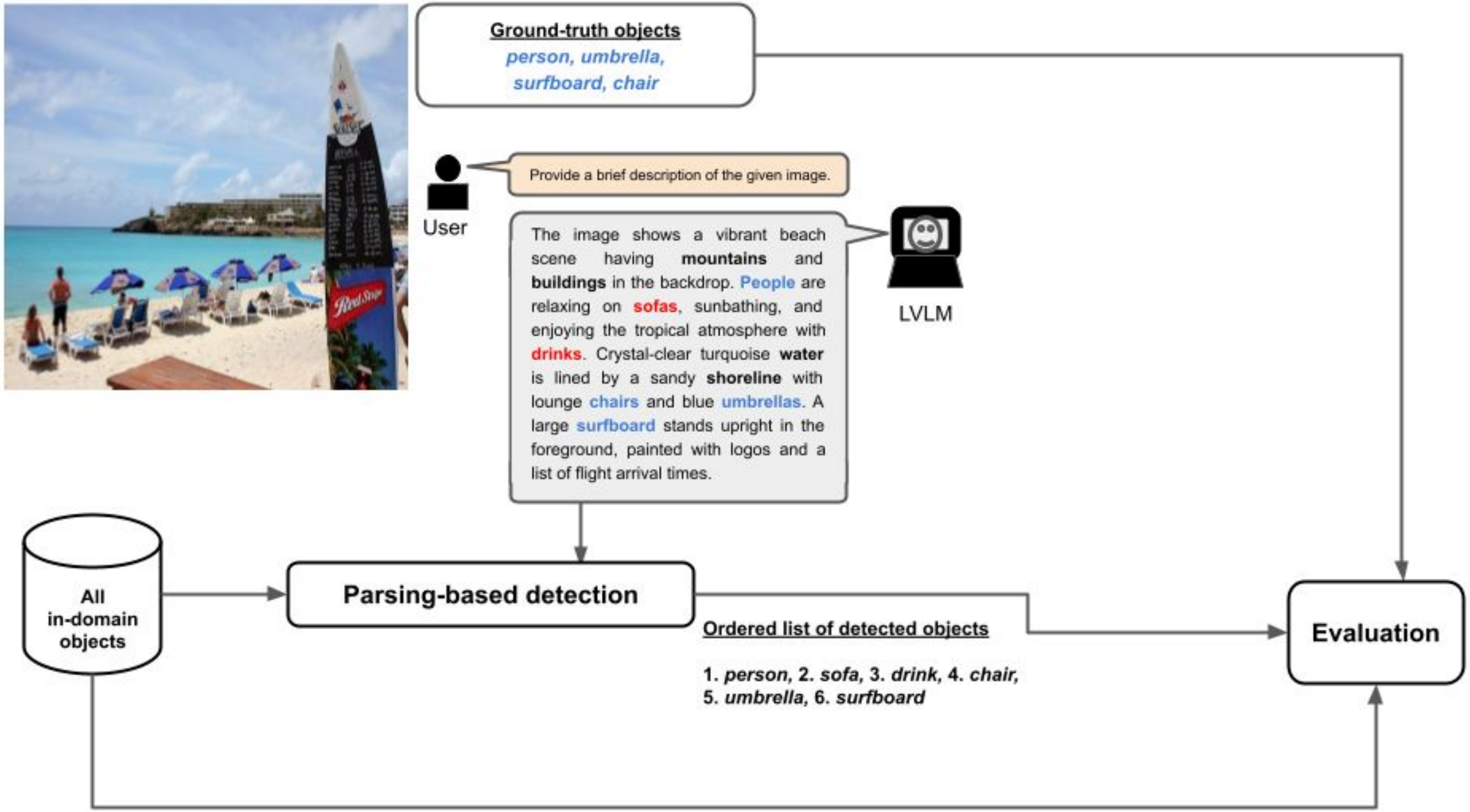
  For LLaVA-7B **20%** of the hallucinated objects co-occurred in the training dataset with at least one preceding objects.

- Can we quantify the semantic influence of ground-truth objects, frequently occurring objects, and past objects on hallucinated objects?

Our contributions

- ***Understanding the Influence of Generation Order:*** We analyze how the sequence of already generated objects affects further hallucinations.
- ***Semantic Analysis of Hallucinations:*** We use word embeddings to examine the relationship between hallucinated objects, ground-truth objects, already generated objects, and frequently occurring objects.
- ***Detecting Out-of-Domain Hallucinations:*** Our approach enhances object detection using LLM-generated captions and verifies the existence of previously unseen objects with an ensemble of LVLMs.

Overview



Ground-truth objects

person, umbrella, surfboard, chair

User

Provide a brief description of the given image.

LVLM

The image shows a vibrant beach scene having **mountains** and **buildings** in the backdrop. People are relaxing on **sofas**, sunbathing, and enjoying the tropical atmosphere with **drinks**. Crystal-clear turquoise **water** is lined by a sandy **shoreline** with lounge chairs and blue umbrellas. A large surfboard stands upright in the foreground, painted with logos and a list of flight arrival times.

All in-domain objects

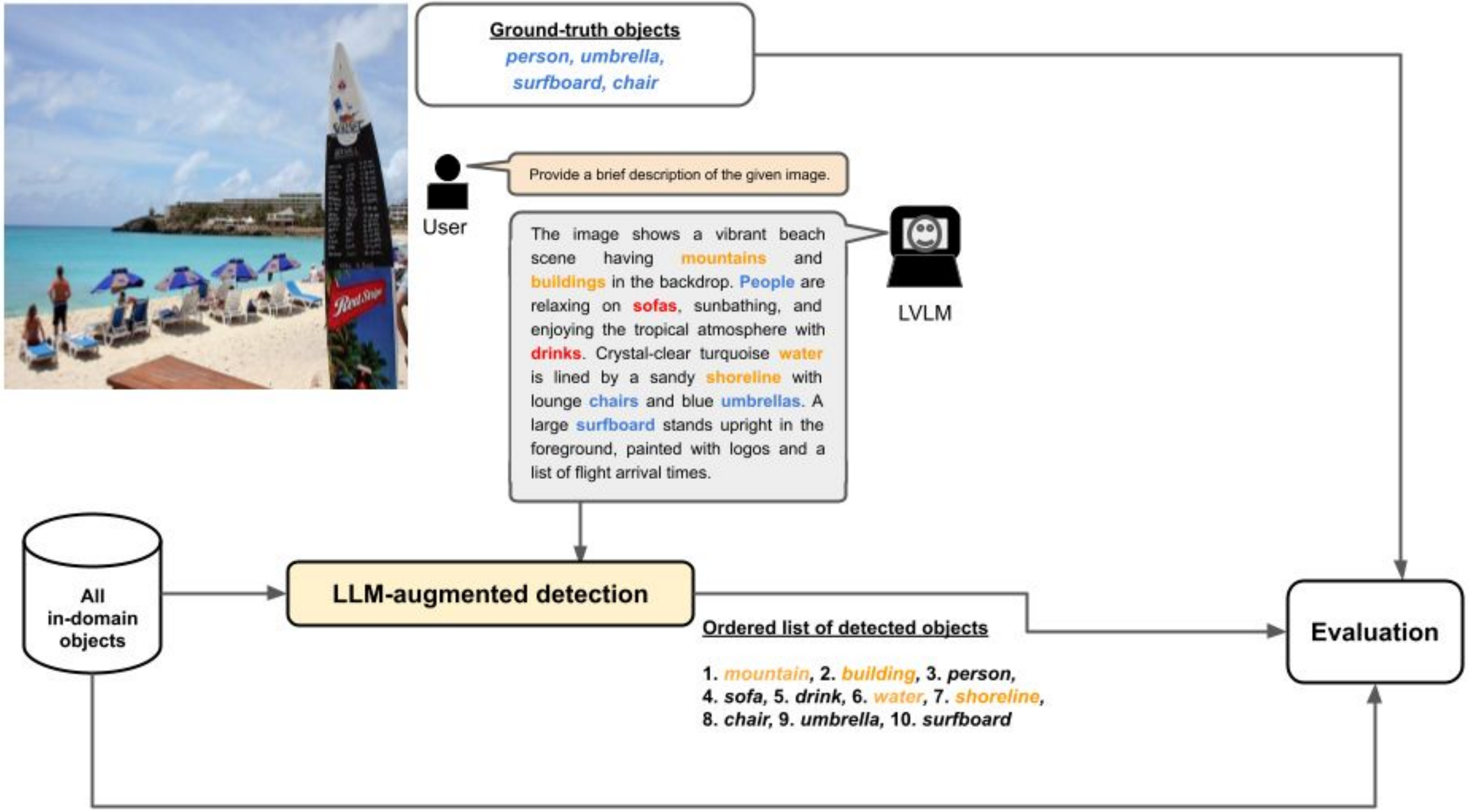
Parsing-based detection

Ordered list of detected objects

1. *person,* 2. *sofa,* 3. *drink,* 4. *chair,* 5. *umbrella,* 6. *surfboard*

Evaluation

Overview



Provide a brief description of the given image.

The image shows a vibrant beach scene having mountains and buildings in the backdrop. People are relaxing on sofas, sunbathing, and enjoying the tropical atmosphere with drinks. Crystal-clear turquoise water is lined by a sandy shoreline with lounge chairs and blue umbrellas. A large surfboard stands upright in the foreground, painted with logos and a list of flight arrival times.

Ordered list of detected objects

1. *mountain*, 2. *building*, 3. *person*, 4. *sofa*, 5. *drink*, 6. *water*, 7. *shoreline*, 8. *chair*, 9. *umbrella*, 10. *surfboard*

LLM-augmented object detection

1. **Detection -** parsing-based objects.

2. **Augmentation -** using LLaMA-2-7B-Chat

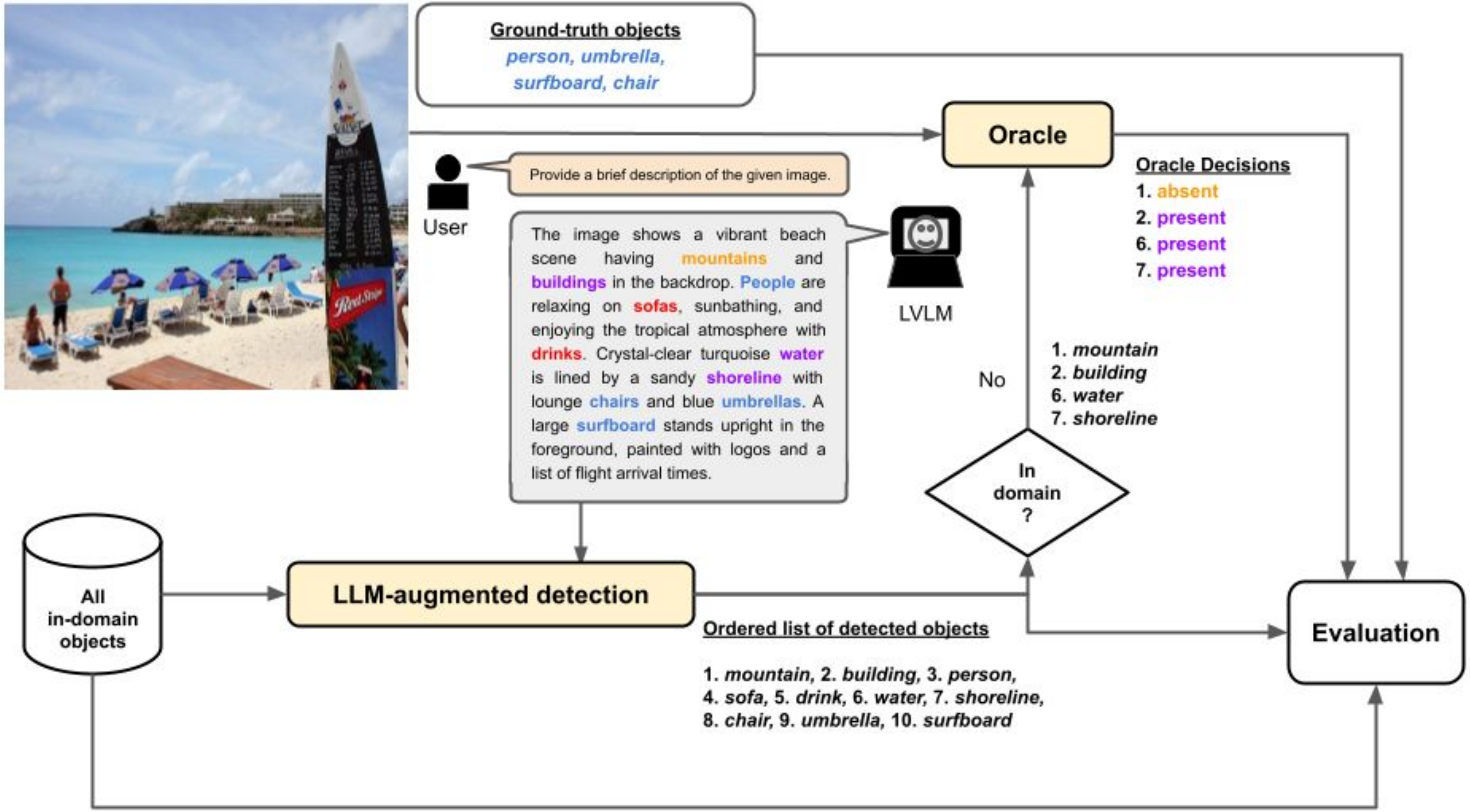
   Few-shot prompt (k=5):
   "**Caption:** The image depicts a group of zebras standing and grazing in a lush, grassy field. There are three zebras in total, with one positioned on the left side of the field, another in the center, and the third on the right side. They are enjoying their time in the green pasture, surrounded by trees which add to the serene atmosphere of the scene.
   Objects: ["zebra", "trees", "field"]
   …

   For the 'Caption' given below, return the 'Objects' as a Python list.
   … "

3. **Filtering -** Remove objects not present in the caption and preserve ordering.

LLM-augmented detection achieves (almost) perfect precision and recall.

Overview



Provide a brief description of the given image.

The image shows a vibrant beach scene having mountains and buildings in the backdrop. People are relaxing on sofas, sunbathing, and enjoying the tropical atmosphere with drinks. Crystal-clear turquoise water is lined by a sandy shoreline with lounge chairs and blue umbrellas. A large surfboard stands upright in the foreground, painted with logos and a list of flight arrival times.

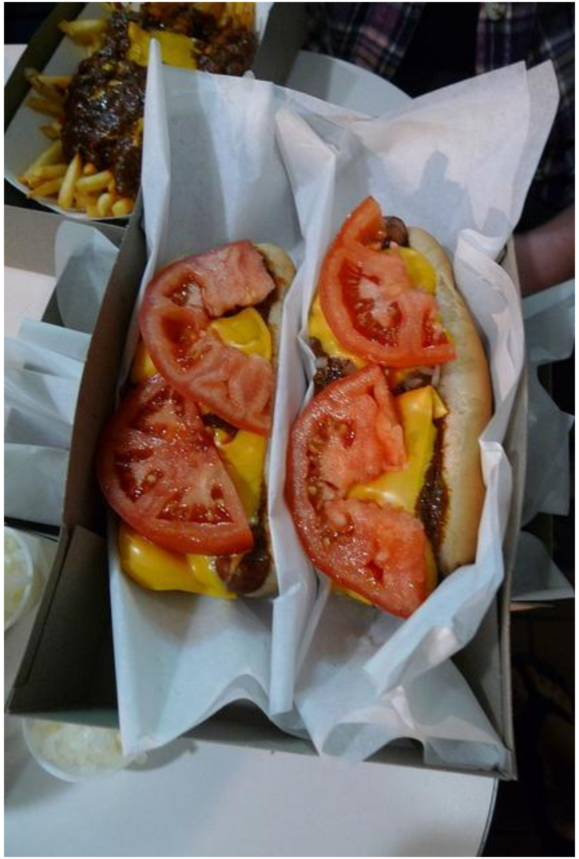
Oracle using an ensemble of LVLMs

Oracle verification: *"Does the image contain <object>? Please respond with only Present or Absent."*

- Majority voting among InstructBLIP, LLaVA, MiniGPT-4, and mPLUG-Owl

- Human evaluation for MultimodalGPT on 100 MSCOCO validation images:

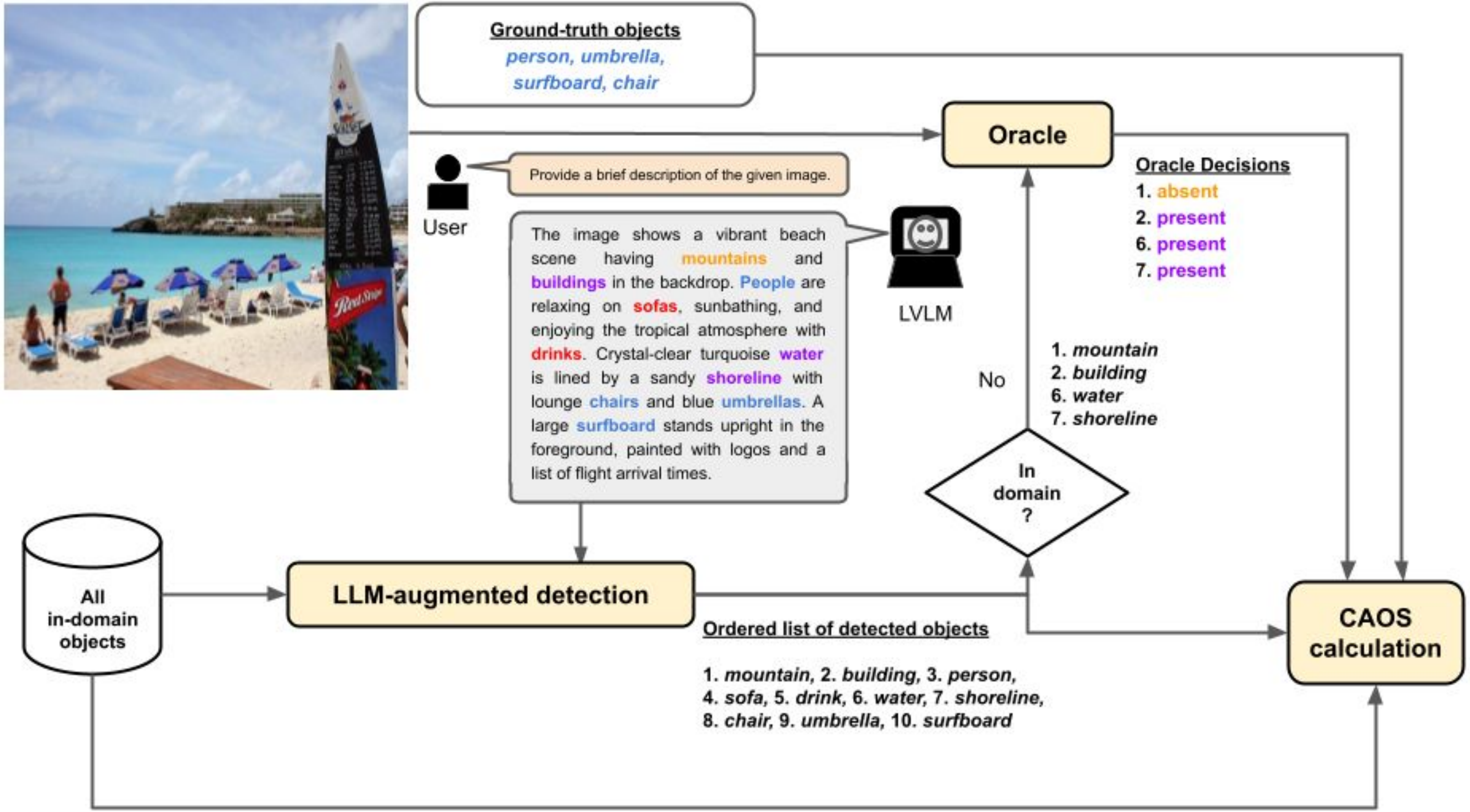
○ InstructBLIP:	**89.57%**
○ LLaVA:	88.42%
○ MiniGPT-4:	84.94%
○ mPLUG-Owl:	84.94%
○ Ensemble:	**93.43%**



MSCOCO objects:
sandwich

Non-MSCOCO objects:
lettuce: absent
tomato: present
onion: absent
ingredients: present

Overview

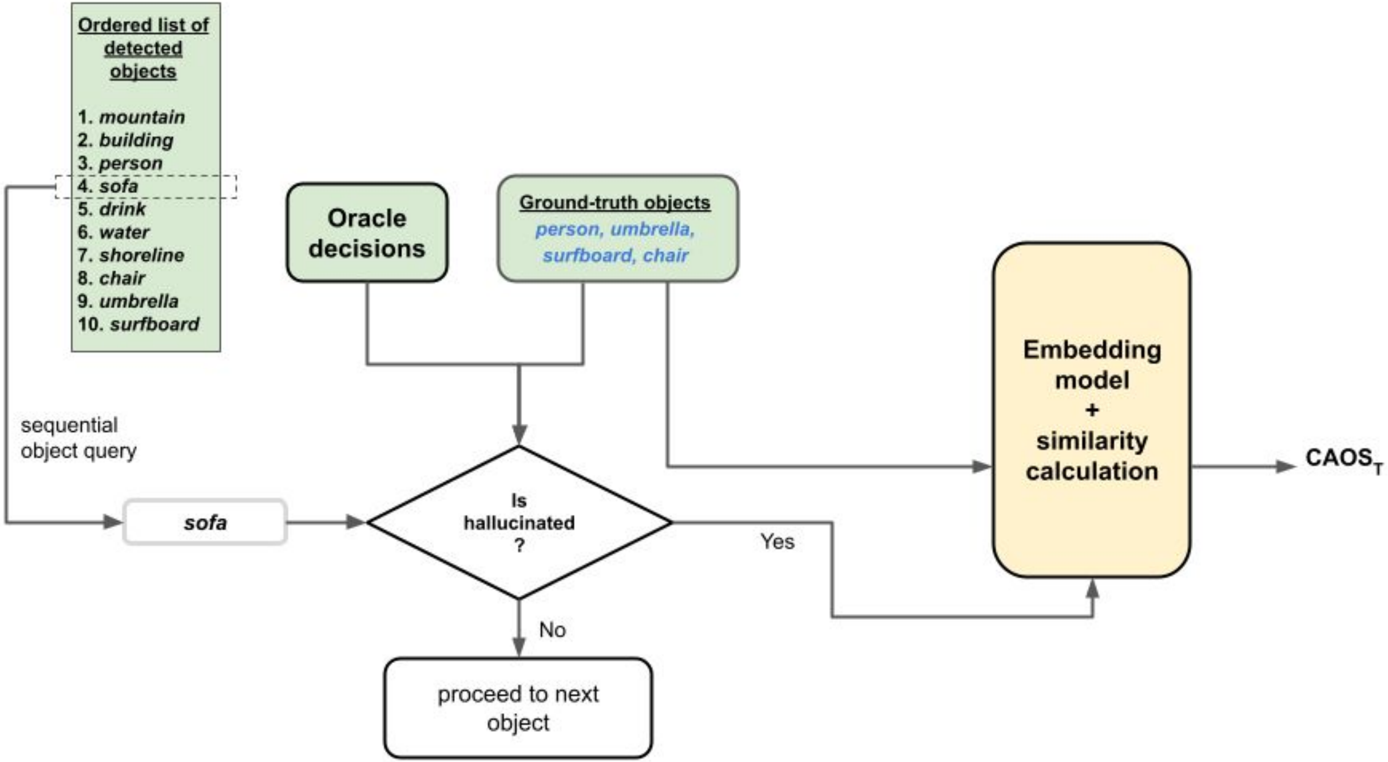


Provide a brief description of the given image.

The image shows a vibrant beach scene having mountains and buildings in the backdrop. People are relaxing on sofas, sunbathing, and enjoying the tropical atmosphere with drinks. Crystal-clear turquoise water is lined by a sandy shoreline with lounge chairs and blue umbrellas. A large surfboard stands upright in the foreground, painted with logos and a list of flight arrival times.

1. mountain, 2. building, 3. person,
4. sofa, 5. drink, 6. water, 7. shoreline,
8. chair, 9. umbrella, 10. surfboard

Similarity calculation

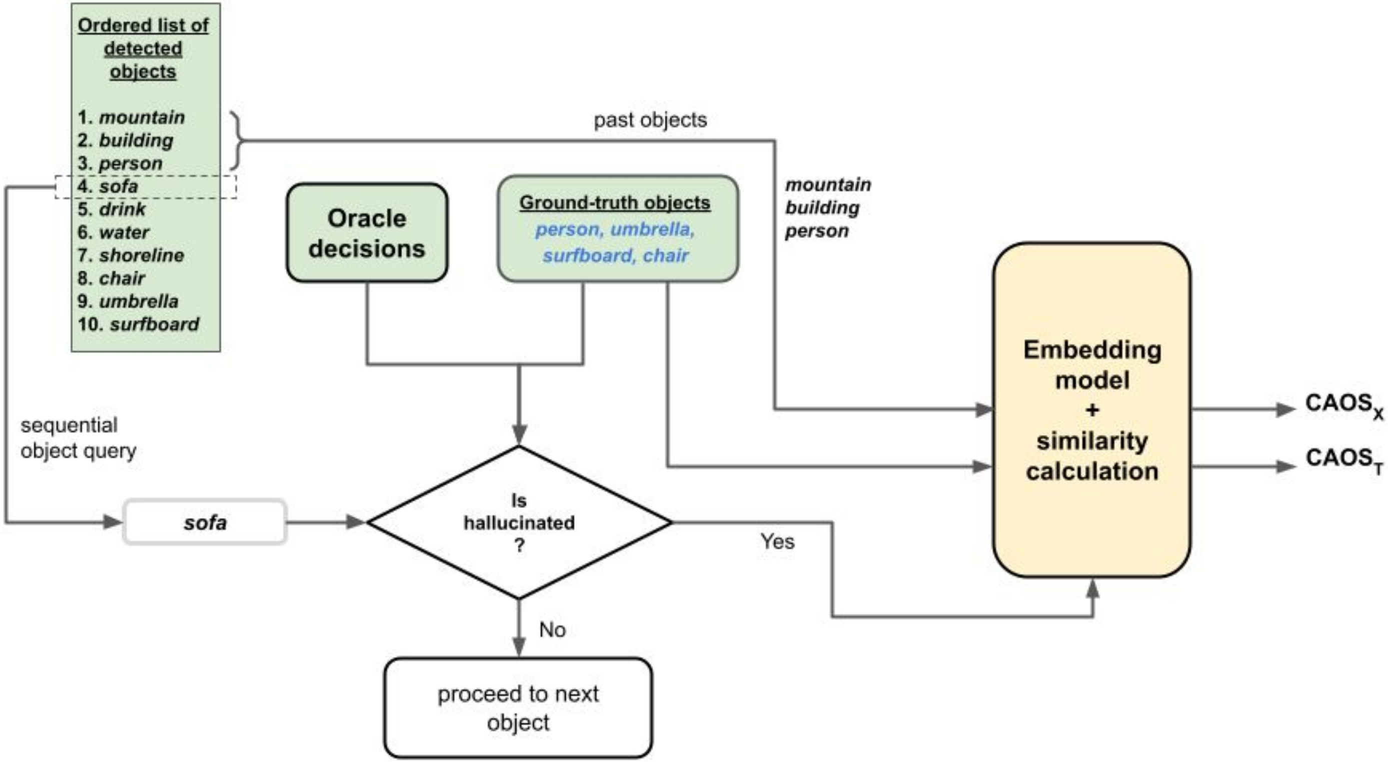


Embedding Models:

- ***GLoVe*** (Pennington et al. 2014) - based on co-occurence
- ***MiniLM-L6*** (Wang et al. 2020) - semantically relations based on a large number of language tasks

Similarity: Cosine Similarity

Similarity calculation

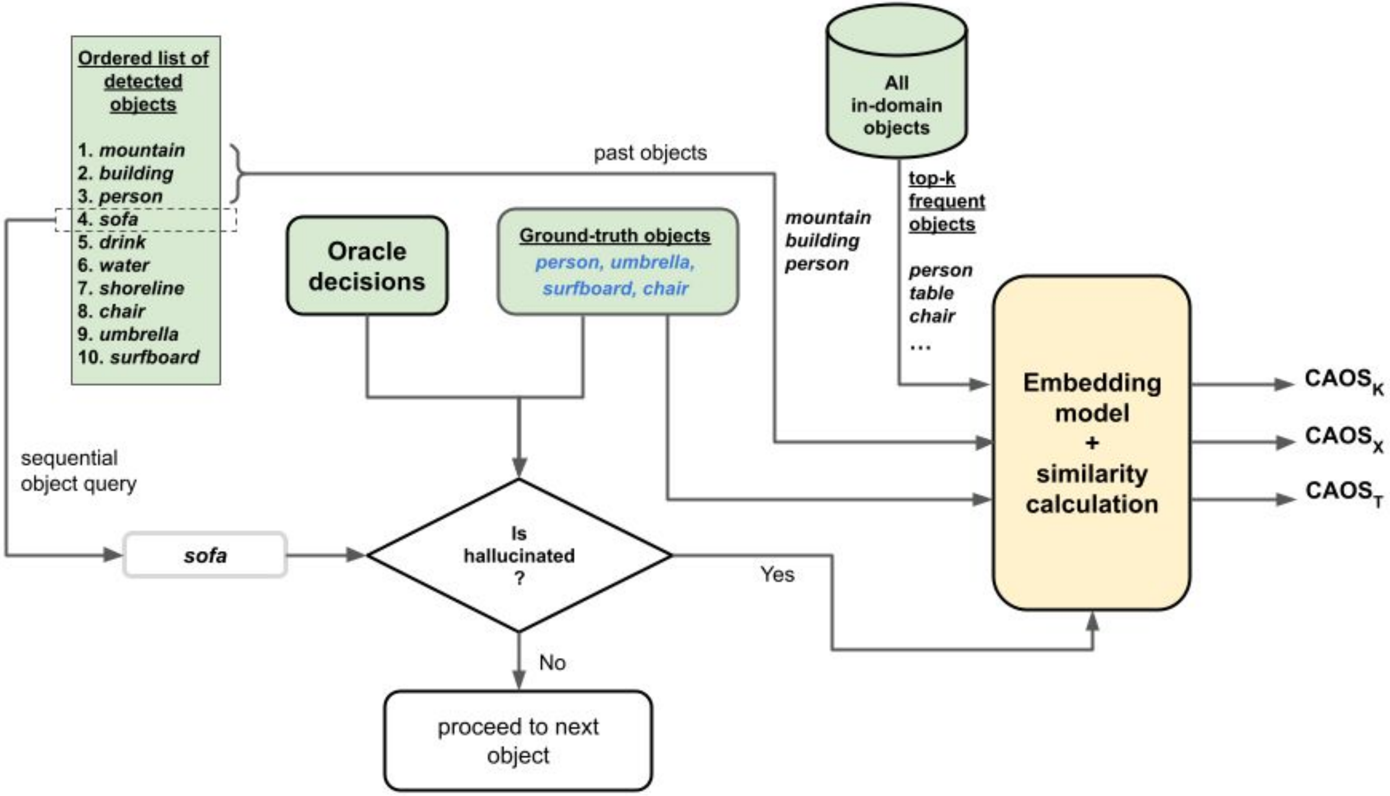


Embedding Models:

- ***GLoVe*** (Pennington et al. 2014) - based on co-occurence
- ***MiniLM-L6*** (Wang et al. 2020) - semantically relations based on a large number of language tasks

Similarity: Cosine Similarity

Similarity calculation

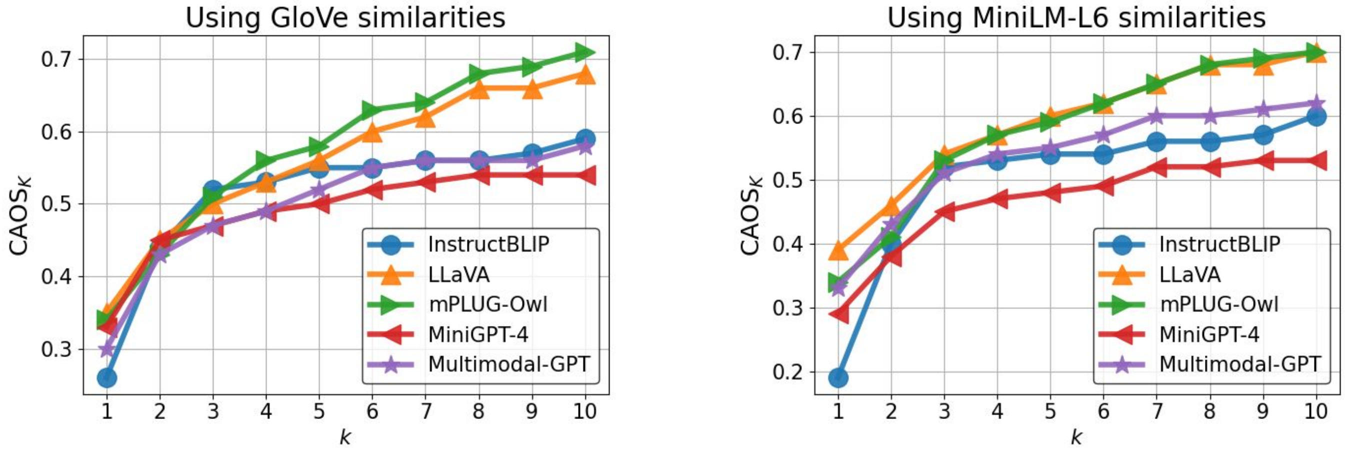


Embedding Models:

- ***GLoVe*** (Pennington et al. 2014) - based on co-occurence
- ***MiniLM-L6*** (Wang et al. 2020) - semantically relations based on a large number of language tasks

Similarity: Cosine Similarity

Knee-point analysis for CAOS_K



- *CAOS_K scores for all models saturate to some extent at k = 3*
- *Top-3 most frequent objects disproportionately appear as hallucinations.*

Context-Aware Object Similarities

CAOS scores:

CAOS_T = Maximum similarity with ground-truth objects

CAOS_X = Maximum similarity with already generated and ground-truth objects

CAOS_K = Maximum similarity with top-K frequently occurring in-domain objects

Context-Aware Object Similarities

CAOS scores:

CAOS_{T} = Maximum similarity with ground-truth objects

CAOS_{X} = Maximum similarity with already generated and ground-truth objects

CAOS_{K} = Maximum similarity with top-K frequently occurring in-domain objects

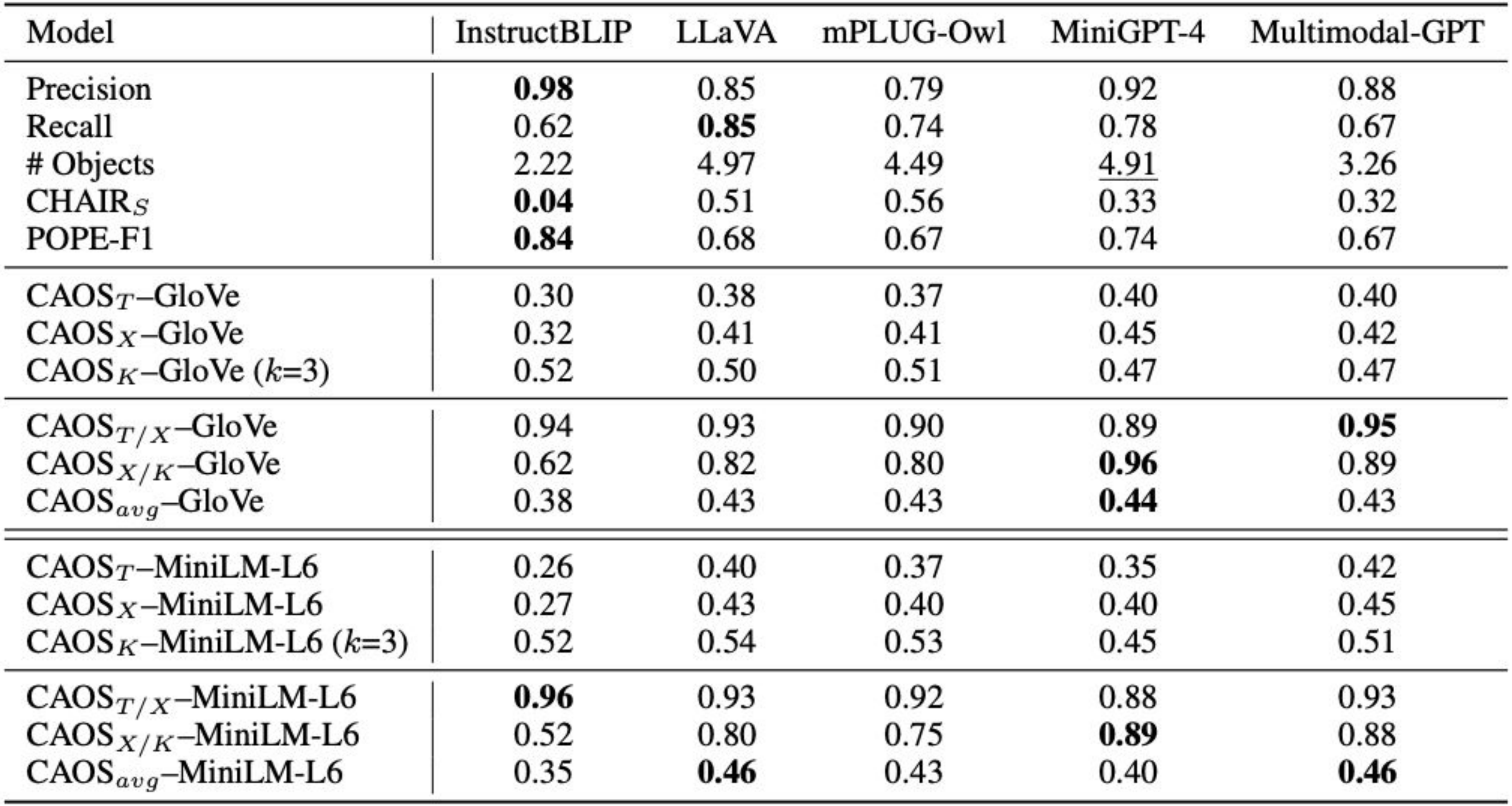
CAOS metrics (Higher is better):

$\mathrm{CAOS}_{T/X} = \mathrm{CAOS}_{T} / \mathrm{CAOS}_{X}$

$\mathrm{CAOS}_{X/K} = \mathrm{CAOS}_{X} / \mathrm{CAOS}_{K}.$

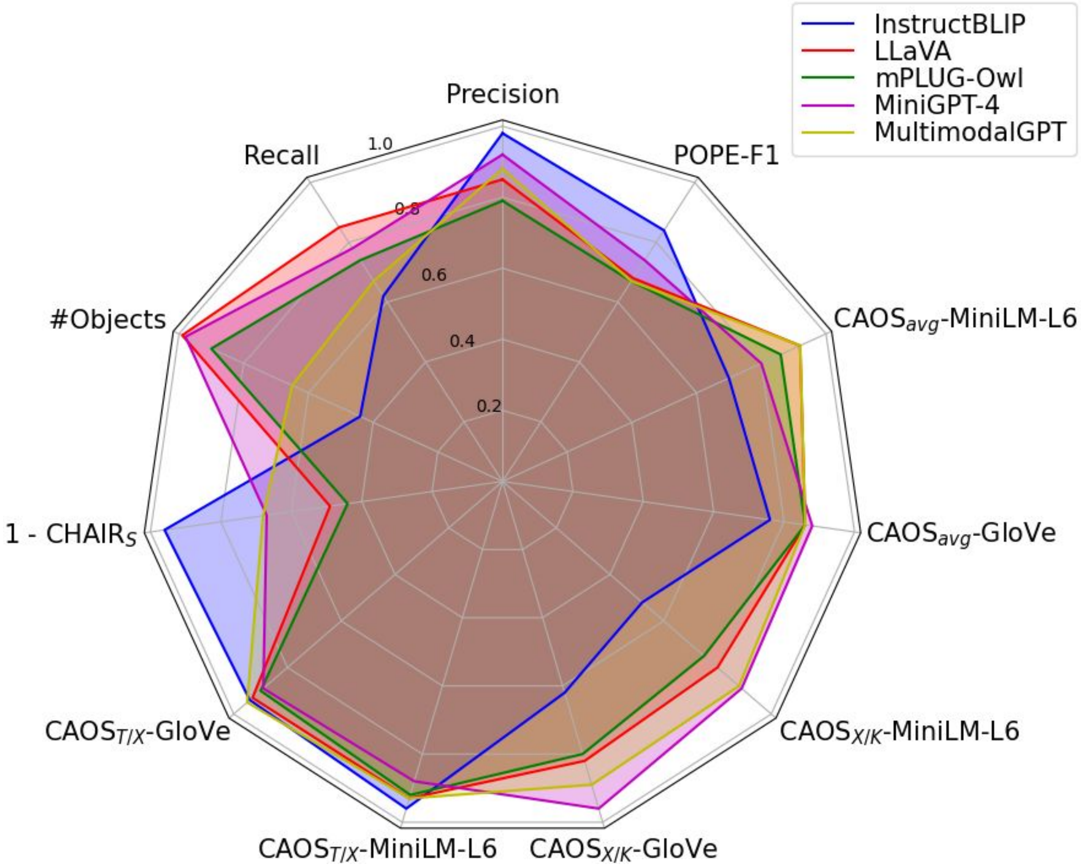
$\mathrm{CAOS}_{\mathrm{avg}} = (\mathrm{CAOS}_{T} + \mathrm{CAOS}_{X} + \mathrm{CAOS}_{K})/3$

CAOS-based comparison of LVLMs

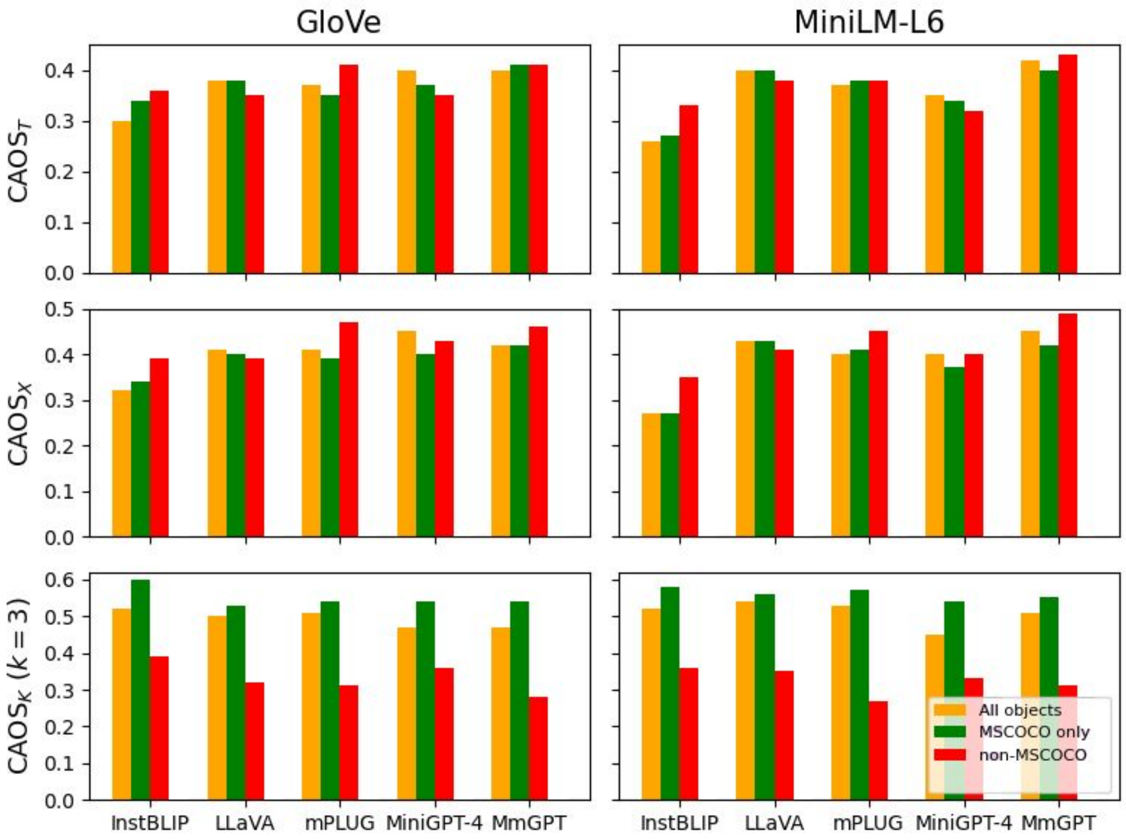


Model	InstructBLIP	LLaVA	mPLUG-Owl	MiniGPT-4	Multimodal-GPT
Precision	**0.98**	0.85	0.79	0.92	0.88
Recall	0.62	**0.85**	0.74	0.78	0.67
# Objects	2.22	4.97	4.49	4.91	3.26
CHAIR$_S$	**0.04**	0.51	0.56	0.33	0.32
POPE-F1	**0.84**	0.68	0.67	0.74	0.67
CAOS$_T$–GloVe	0.30	0.38	0.37	0.40	0.40
CAOS$_X$–GloVe	0.32	0.41	0.41	0.45	0.42
CAOS$_K$–GloVe (k=3)	0.52	0.50	0.51	0.47	0.47
CAOS$_{T/X}$–GloVe	0.94	0.93	0.90	0.89	**0.95**
CAOS$_{X/K}$–GloVe	0.62	0.82	0.80	**0.96**	0.89
CAOS$_{avg}$–GloVe	0.38	0.43	0.43	**0.44**	0.43
CAOS$_T$–MiniLM-L6	0.26	0.40	0.37	0.35	0.42
CAOS$_X$–MiniLM-L6	0.27	0.43	0.40	0.40	0.45
CAOS$_K$–MiniLM-L6 (k=3)	0.52	0.54	0.53	0.45	0.51
CAOS$_{T/X}$–MiniLM-L6	**0.96**	0.93	0.92	0.88	0.93
CAOS$_{X/K}$–MiniLM-L6	0.52	0.80	0.75	**0.89**	0.88
CAOS$_{avg}$–MiniLM-L6	0.35	**0.46**	0.43	0.40	**0.46**

CAOS-based comparison of LVLMs

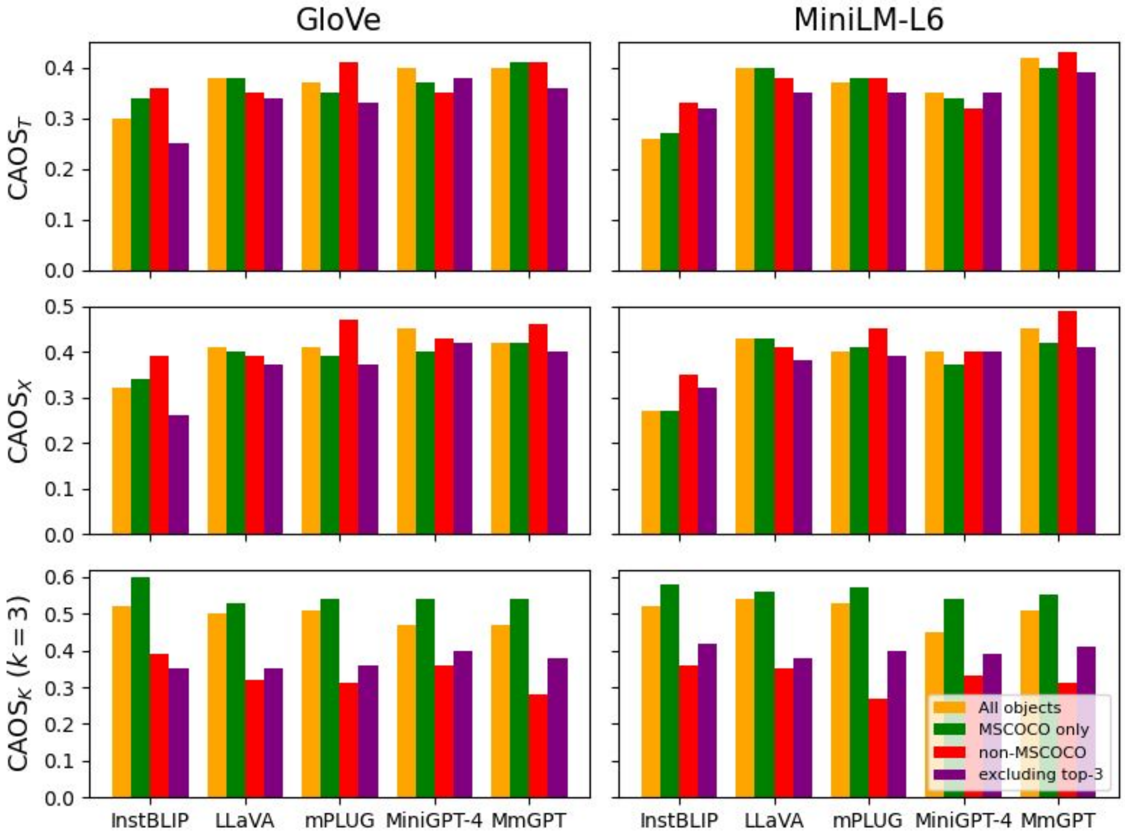


Comparison across various groups



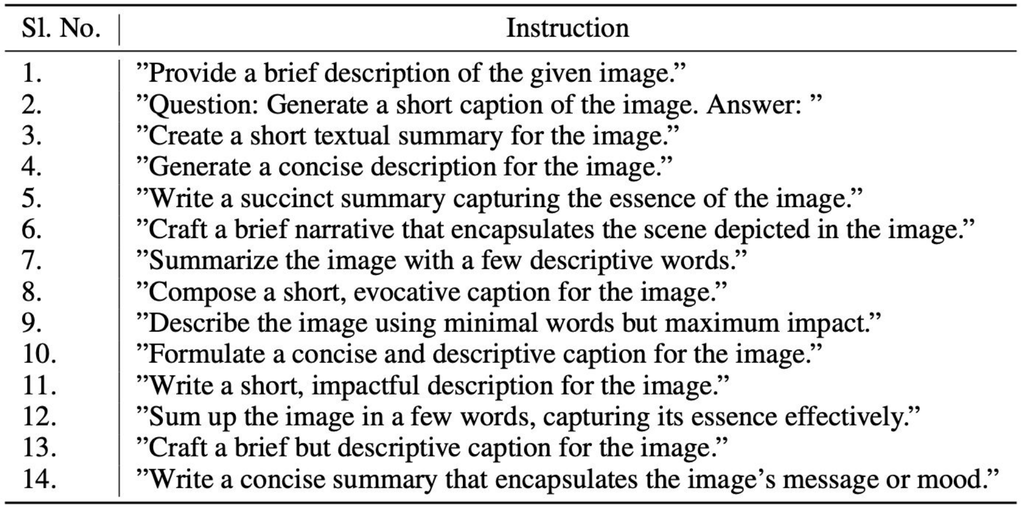
- CAOS scores are largely similar for in-domain and out-of-domain hallucinations.
- Out-of-domain hallucinations are still influenced by frequent MSCOCO objects, though to a lesser extent.

Comparison across various groups



- CAOS scores are largely similar for in-domain and out-of-domain hallucinations.
- Out-of-domain hallucinations are still influenced by frequent MSCOCO objects, though to a lesser extent.
- Excluding the top-3 most frequent MSCOCO objects leads to a dip in CAOS_K scores, indicating that these objects disproportionately appear as hallucinations across models.

Sensitivity to diverse prompts

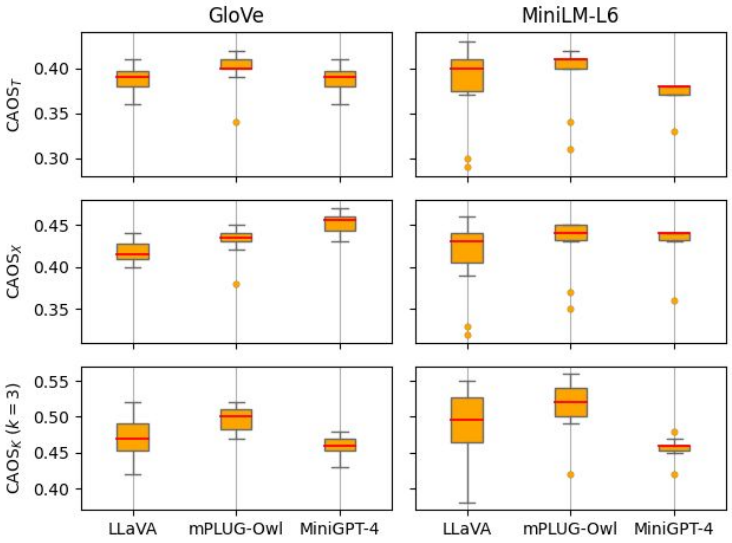


Sl. No.	Instruction
1.	”Provide a brief description of the given image.”
2.	”Question: Generate a short caption of the image. Answer: ”
3.	”Create a short textual summary for the image.”
4.	”Generate a concise description for the image.”
5.	”Write a succinct summary capturing the essence of the image.”
6.	”Craft a brief narrative that encapsulates the scene depicted in the image.”
7.	”Summarize the image with a few descriptive words.”
8.	”Compose a short, evocative caption for the image.”
9.	”Describe the image using minimal words but maximum impact.”
10.	”Formulate a concise and descriptive caption for the image.”
11.	”Write a short, impactful description for the image.”
12.	”Sum up the image in a few words, capturing its essence effectively.”
13.	”Craft a brief but descriptive caption for the image.”
14.	”Write a concise summary that encapsulates the image’s message or mood.”

Table 3: List of all instructions used for our experiments.

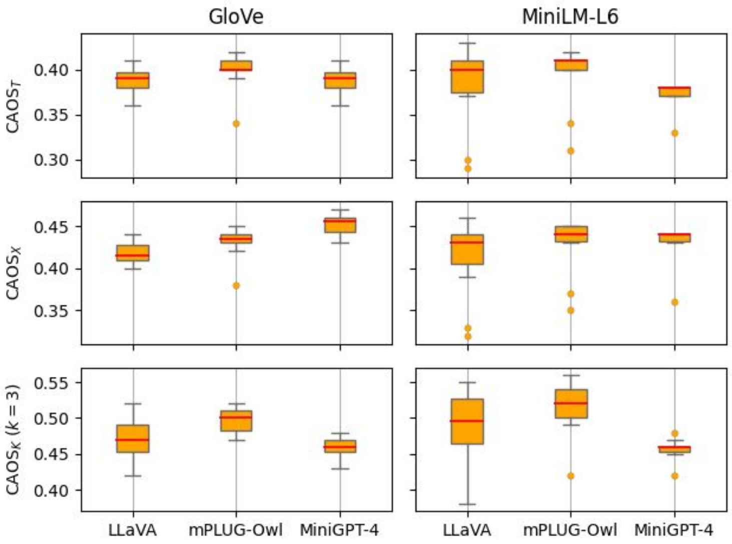
Sensitivity to diverse prompts

Results



- CAOS scores are largely stable to changes in instructions.

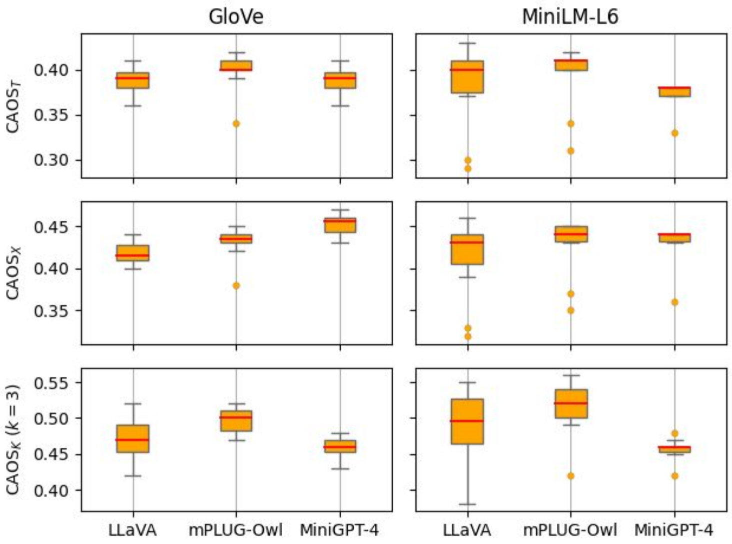
Sensitivity to diverse prompts



- CAOS scores are largely stable to changes in instructions.

- CAOS scores have different ranges across the different LVLMs.

Sensitivity to diverse prompts



- CAOS scores are largely stable to changes in instructions.
- CAOS scores have different ranges across the different LVLMs.
- CAOS scores calculated using MiniLM-L6 embeddings seem to be slightly more prone to having outliers than their corresponding GloVe counterparts.

References

- Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. In Empirical Methods in Natural Language Processing (EMNLP).

- Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating object hallucination in large vision language models. arXiv preprint arXiv:2305.10355.

- Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

- Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. Minilm: Deep self-attention distillation for taskagnostic compression of pre-trained transformers. Advances in Neural Information Processing Systems, 33: 5776–5788.

Thank You!

Please check-out our paper for more details